TI 2021-035/V
Tinbergen Institute Discussion Paper

Disparities in socio-economic status and BMI in the UK are partly due to genetic and environmental luck

Casper A.P. Burik[1]
Hyeokmoon Kweon[1]
Philipp D. Koellinger[2,1]

[1] Department of Economics, School of Economics and Business, Vrije Universiteit Amsterdam
[2] La Follette School of Public Affairs, University of Wisconsin-Madison

Disparities in socio-economic status and BMI in the UK are partly due to genetic and environmental luck

Casper A.P. Burik[1], Hyeokmoon Kweon[1] and Philipp D. Koellinger*[2,1]

1. Department of Economics, School of Economics and Business, Vrije Universiteit Amsterdam
2. La Follette School of Public Affairs, University of Wisconsin-Madison
* Corresponding author: koellinger@wisc.edu

Abstract

Two family-specific lotteries take place during conception— a social lottery that determines who our parents are and which environment we grow up in, and a genetic lottery that determines which part of their genomes our parents pass on to us. The outcomes of these lotteries create inequalities of opportunity that can translate into disparities in health and socioeconomic status. Here, we estimate a lower bound for the relevance of these two lotteries for differences in education, income and body mass index in a sample of 38,698 siblings in the UK who were born between 1937 and 1970. Our estimates are based on models that combine family-specific effects with gene-by-environment interactions. We find that the random differences between siblings in their genetic endowments clearly contribute towards inequalities in the outcomes we study. Our rough proxy of the environment people grew up in, which we derived from their place of birth, are also predictive of the studied outcomes, but not beyond the relevance of family environment. Our estimates suggest that at least 13 to 17 percent of the inequalities in education, wages and BMI in the UK are due to inequalities in opportunity that arise from the outcomes of the social and the genetic lottery.

1. Introduction

It has long been recognized that parent's health and socio-economic status (SES) are strong predictors for their children's health, educational attainment and income later in life. Furthermore, health, educational attainment, and income are all heritable to some degree (Benjamin et al., 2012; de Vlaming et al., 2017; Polderman et al., 2015; Taubman, 1976). Thus, parents do not only influence their children via the rearing environment they provide for them, but also by the random combination of the genes they pass on to their offspring. This creates two major sources of differences in opportunities at conception in the form of exogenously determined environmental and genetic endowments. Disparities in important life outcomes that arise from differences in opportunity are often viewed as unfair and less desirable than inequality that is created by active choices and agency (e.g. due to hard work). This may have policy implications because people tend to favour redistribution policies more when inequalities in opportunity and luck are major drivers of inequality (Alesina & La Ferrara, 2005; Alesina, Stantcheva, & Teso, 2018; Almås, Cappelen, Sørensen, & Tungodden, 2010; Cappelen, Konow, Sørensen, & Tungodden, 2013; Clark & D'Ambrosio, 2015; Gromet, Hartson, & Sherman, 2015). Thus, studying the relative importance of inequalities of opportunity for important life outcomes is of fundamental importance to discussions about fairness and policy.

Genetic factors that are linked to socio-economic status are a reflection of social realities. For example, societies that value high cognitive performance in schools and labour markets will tend to exhibit that genetic factors that are linked with cognitive health are also related to socio-economic outcomes such as educational attainment or income. Thus, genes do not operate in a vacuum – their effects are partially contingent on environmental factors. Furthermore, specific environments and genes may also interact with each other (Barcellos, Carvalho, & Turley, 2018, 2020; Schmitz & Conley, 2017a, 2017b), potentially further exacerbating the importance of the genetic and the social lottery as a source of inequality.

Recent advances in genetics have made it possible to measure genetic differences between people comprehensively, providing researchers with new opportunities to study the potential relevance of genetic luck and to investigate how exogenously given genetic and environmental endowments can interact to cause inequalities (Harden & Koellinger, 2020). Moreover, increases in sample size have led to publicly available summary statistics from large-scale genome-wide association studies (GWAS) for many outcomes related to SES and health, such as educational attainment (Lee et al., 2018), household income (Hill et al., 2019), occupational wages (Kweon et al., 2020), body fat percentage (Lu et al., 2016), and body mass index (BMI) (Locke et al., 2015). The estimated effects of these GWAS can be summarized in linear indices that are called polygenic indices (PGI[a]) (Daetwyler, Villanueva, & Woolliams, 2008; Dudbridge, 2013). Although PGI capture only a part of the heritability of a trait because they are measured with error (Daetwyler et al., 2008; DiPrete, Burik, & Koellinger, 2018; Dudbridge, 2013), they nevertheless provide a valuable new tool to analyse genetic contributions to inequality and to study potential interactions between genetic endowments and specific environmental conditions (Barcellos et al., 2018; Harden & Koellinger, 2020).

The goal of this study is to estimate a lower bound for the relevance of environmental and genetic luck and their interactions for important life outcomes. We employ data from the UK Biobank, which is currently the largest publicly available sample of genotyped siblings in the world (38,698 individuals). Genetic differences between biological siblings are due to the natural experiment of meiosis. During meiosis, the two copies of each parental chromosome are randomly combined and then separated to create a set of two gametes (e.g., two eggs or two sperm), each of which contains only one new, resampled copy of each chromosome. The resulting genetic differences between full siblings are therefore random and independent from family-specific ancestry and environmental factors that vary between families.

Our choice of outcome variables was specified in a pre-registered analysis plan[b] and driven by considerations about data availability and statistical power. In the socioeconomic domain, we focus on educational attainment (EA) and hourly wages. Both are key components of socio-economic status, and both are linked to happiness (Boyce, Brown, & Moore, 2010; Frijters, Haisken-DeNew, & Shields, 2004), health, and longevity (Adler & Rehkopf, 2008; Stringhini et al., 2017; Wilkinson & Marmot, 2003). In the health domain, we focus on BMI as a proxy for morbidity that is also linked to

[a] Here we follow the recent change proposed in Becker et al. (2021) from polygenic (risk) score to polygenic index to make it less likely to be wrongly interpreted as a value judgement.

[b] Our pre-registered analysis plan can be accessed here: https://osf.io/wf56h/

mortality (Mokdad et al., 2003) and many other health outcomes. Importantly, for all three outcomes, large-scale GWAS results are available that allow constructing PGI that capture a substantial part of the heritability of these traits (Kweon et al., 2020; Lee et al., 2018; Locke et al., 2015).

We extracted measures of potentially relevant environmental factors during early childhood from available information about place of birth. Chetty and Hendren (2018a, 2018b) show childhood neighbourhood affects later-life outcomes like educational attainment and income. Amongst other factors, they find that school quality has a positive effect. Furthermore, neighbourhood SES has been shown to be related to infant health and infant mortality rate in the UK (Weightman et al., 2012). In this study, we used the local average school leaving age and the district mortality rate at the place of birth as measures of childhood environment.

Importantly, our genetic and environmental variables only capture a part of the ways in which the outcomes of the genetic and the social lottery may influence outcomes later in life, and all our variables are subject to substantial measurement error, which attenuates the estimated effects of these two lotteries towards zero. Thus, our study estimates a conservative lower bound for the potential relevance of these two sources of luck on lifetime outcomes.

In addition to the linear effects of PGI and childhood environments, we also investigate potential interaction effects between them. Numerous studies have begun identifying relevant gene-by-environment interactions both on SES and health outcomes. One example of a study on inequality and gene-by-environment interaction is Belsky et al. (2018), who study social mobility in several cohorts using a PGI based on GWAS results for educational attainment (EA) from Lee et al. (2018). They find that both parental SES and the genetic endowment of the child contribute to social mobility. In analyses that control for family fixed effects, the sibling with the higher PGI for EA is found to be more likely to have higher SES later in life, suggesting that random genetic differences between siblings contribute towards social mobility. While Belsky et al. also investigate gene-by-environment interactions and conduct analyses within-families, they do not combine the two approaches. This makes their results of the gene-by-environment analyses more difficult to interpret because they may be confounded by unobserved family-specific environments that correlated with genetic endowments (Harden & Koellinger, 2020; Schmitz & Conley, 2017a).

Similar to the study of Belsky et al. (2018), many gene-by-environment studies are difficult to interpret due to sensitivity to confounding from unobserved family-specific environments and population structure that are correlated with both the environmental measure and the underlying genetic factors (Harden & Koellinger, 2020; Schmitz & Conley, 2017a).

One of the solutions proposed in the literature is the use of natural experiments (Schmitz & Conley, 2017a), for instance using policy interventions (Schmitz & Conley, 2017b). Barcellos et al. (2018) take this approach and study the effects of genes and education on health outcomes in the UK Biobank. They make use of a well-known compulsory schooling age reform in the United Kingdom in in 1972 as a quasi-experiment and find that an increase in education can reduce health differences related to genetic risk of obesity. Furthermore, Barcellos et al. (2020) use a similar approach to Barcellos et al. (2018) to investigate the effects of the same schooling reform on education and wages later in life as well as the interaction between birth place effects and PGI. They found that the schooling reform reduced differences in educational attainment across birth places, but benefitting those with high PGI for EA the most. The effect of education on wage was twice as high in the top tercile of the PGI, compared to the bottom and middle terciles. While policy reforms that induce as-good-as-random variation in education are a common method to identify causal effects, the results are often specific to the policy and context that is being studied (Rosenzweig & Wolpin, 2000).

Our study investigates the effects of environmental and genetic luck and their possible interactions for important life outcomes using a novel approach. We combine measures of early childhood environment with random genetic differences between siblings in a within-family design. The random genetic differences between siblings are by definition independent from shared environments that are not captured by our early life exposures of interest, thereby circumventing the endogeneity problem that most gene-environment studies suffer from. Furthermore, we investigate different gene-environment interactions than those investigated in earlier work.

2. Materials

Sample

The UK Biobank is a large population-based longitudinal study, designed to study health in middle aged and older UK citizens (Fry et al., 2017; Sudlow et al., 2015). The participants were between 40 and 69 years old when they entered the study between 2006 and 2010. Participants answered a wide array of survey questions about their life and health and various physical measurements and biological samples (saliva, blood and urine) were taken during an assessment centre visit. Almost all participants were genotyped and all participants gave broad consent for research related to health and well-being. We restrict our analyses to individuals of European descent to limit possible confounding due to population structure. Identification of European ancestry was done by the UK Biobank based on principal component analysis with the 1000 Genomes project reference panel (1000 Genomes Project Consortium et al., 2015).

Early Childhood Environment

The UK Biobank does not contain direct measures of early childhood environment that are pertinent to our research question. To obtain proxies of socio-economic environment during childhood, we used birth place coordinates. We matched these coordinates to district-level early-childhood exposures that we obtained from historical data made available by Vision of Britain (Southall, 2011)[c]. Specifically, we use the local average school leaving age and the infant mortality rate at the district level (see SI section 2 for details).

Outcomes

Following prior literature, we measured educational attainment in years of schooling (see SI section 5). To obtain a proxy for individual income, we imputed occupational wages from standardized

[c] www.visionofbritain.org.uk

occupation codes using an algorithm developed by Kweon et al. (2020). The imputed values reflect the logarithm of the typical wage per hour for each occupation, adjusted for demographic characteristics such as sex and age. The imputation algorithm utilizes wage data provided by the UK Office of National Statistics, using the British Household Panel Survey to estimate model parameters and the Labour Force Survey for external validation. This procedure primarily captures wage differences between occupations and captures $R^2 \approx 0.50$ of the total variance of hourly wages. Finally, body mass index (BMI) -- our proxy for health -- is based on physical measures taken in the UK Biobank assessment centre.

Polygenic Indices

We constructed PGI using the results of the largest GWAS publicly available for educational attainment, occupational wages, and BMI, which are Lee et al. (2018), Kweon et al. (2020), and Locke et al. (2015) respectively. Furthermore, we used multi-trait analysis of genome-wide association summary statistics (MTAG; Turley et al., 2018) to increase the accuracy of the PGI by including summary statistics of genetically correlated traits. To account for linkage-disequilibrium, we constructed PGI using LDpred (Vilhjálmsson et al., 2015). SI section 3 provides further detail.

3. Methods

First, we mapped the relationships of the outcomes in adulthood with early childhood environment and genetic endowments. We divided the sample into different terciles of the early childhood environment and PGI distributions. We then compared the means of our outcome variables across terciles to visualize how SES and BMI differ based on place of birth and genetic endowments.

We then regressed our outcomes on the PGI, dummy variables for the district terciles, and interaction terms between the two as well as other control variables:

$$y_i = \beta_0 + \beta_1 G_i + \boldsymbol{D_i}\boldsymbol{\beta_2} + (G_i \times \boldsymbol{D_i})\boldsymbol{\beta_3} + \boldsymbol{PC_i}\boldsymbol{\gamma} + \boldsymbol{Z_i}\boldsymbol{\delta} + \epsilon_i \quad (1)$$

where y_i is the outcome of individual i (educational attainment, imputed log hourly wage or BMI), G_i is the PGI for the respective outcome, $\boldsymbol{D_i}$ is a vector with two dummy variables for the middle and top terciles of the distribution of our environmental variable (local average school leaving age or local infant mortality rate), $\boldsymbol{PC_i}$ is a vector of principle components of the genetic data to control for population stratification, $\boldsymbol{Z_i}$ is a vector of other control variables (including year of birth, year of birth squared, year of birth cubed, gender, gender interacted with the year of birth variables and genotyping batch), and ϵ_i is the error term. It should be noted that while our primary interest lies in the estimates for β_1, $\boldsymbol{\beta_2}$ and $\boldsymbol{\beta_3}$, the covariates included in $\boldsymbol{Z_i}$ are also the result of luck in the sense that nobody has an influence of their time and place of birth or their biological sex, either. Thus, all of the variance explained by model (1) can be attributed to luck defined as exogenously given resources that are outside of one’s control.

Next, we re-estimate equation (1) adding family fixed effects. By using family fixed effects, we utilize the random genetic differences between siblings that are by definition independent from shared environments that are not captured by our early life exposures of interest, thereby circumventing endogeneity problems caused by inadequately controlling for unobserved gene-environment correlations. Since our environmental exposures are local-level measures at the birth location, it is plausible that these exposures are not dependent on own genetic effects but only on parental genetic effects and pre-birth family characteristics, which are captured by the family fixed effects. Therefore, these family fixed effects also adjust for potential biases in the estimates of β_1 that result from indirect genetic effects such as genetic nurture (Kong et al., 2018) or any population stratification that is not captured by the principal components of the genetic data.

While the family fixed effects capture many possible biases, it can fail to deliver a within-family estimator for the interaction term, as the interaction term is not guaranteed to be independent of between-family variation (Giesselmann & Schmidt-Catran, 2018, 2020). Therefore, we extend our analyses to control for those sources of between-family variation by adding additional control variables for between-family variation to a random effects model based on Mundlak's work (Mundlak, 1978), extended to account for interaction terms:

$$y_{ij} = \beta_1 G_{ij} + \beta_2 E_{ij} + \beta_3(G_{ij} \times E_{ij}) + \theta_1 \overline{G}_J + \theta_2 \overline{E}_J + \theta_3\left(\overline{G}_J \times E_{ij}\right) + \theta_4\left(G_{ij} \times \overline{E}_J\right) + \boldsymbol{Z}'_{ij}\boldsymbol{\delta_1} + \overline{\boldsymbol{Z}}'_j\boldsymbol{\delta_2} + (u_j + \varepsilon_{ij}) \qquad (2)$$

where $\overline{X}_J$ indicates the family-specific mean of the variable X_{ij} and $(u_j + \varepsilon_{ij})$ is the error component, with u_j as the family-level random effect. All other variables are defined as above. Every variable in this regression was mean-centred, so that the estimated coefficients of the model provide the effect size at the means of all variables.

Estimating this model as a random-effects framework gives within-family estimate for β_1 , β_2, and β_3. The key components of this model are $\overline{G}_J$, $\overline{E}_J$, $\left(\overline{G}_J \times E_{ij}\right)$, and $\left(G_{ij} \times \overline{E}_J\right)$ which control for the unobserved between-family differences in the PGI, the environment measure, and the gene-environment interaction; thereby yielding within-family estimates for β_1 , β_2, and β_3. The within-family means are designed to capture more dimensions of the between family variation than the family fixed effects model. Therefore, a within-family estimate for the gene-environment interaction of this model will not represent a spurious gene-environment interaction (Giesselmann & Schmidt-Catran, 2018, 2020). While the model accounts for many possible sources of bias by including family specific effects and other control variables, it cannot control for all possible sources of omitted variable bias, especially those due to potential indirect genetic effects from siblings on each other, which may limit the causal interpretation of our estimates. However, indirect genetic effects from siblings would lead to a bias towards zero for the estimated effect of the PGI and the interaction term due to possible spill-over effects from the sibling with the higher PGI to the sibling with the lower PGI, decreasing the within-sibling differences in outcomes. Therefore, if sibling effects are present, our estimates are likely to underestimate the direct genetic effects.

4. Results

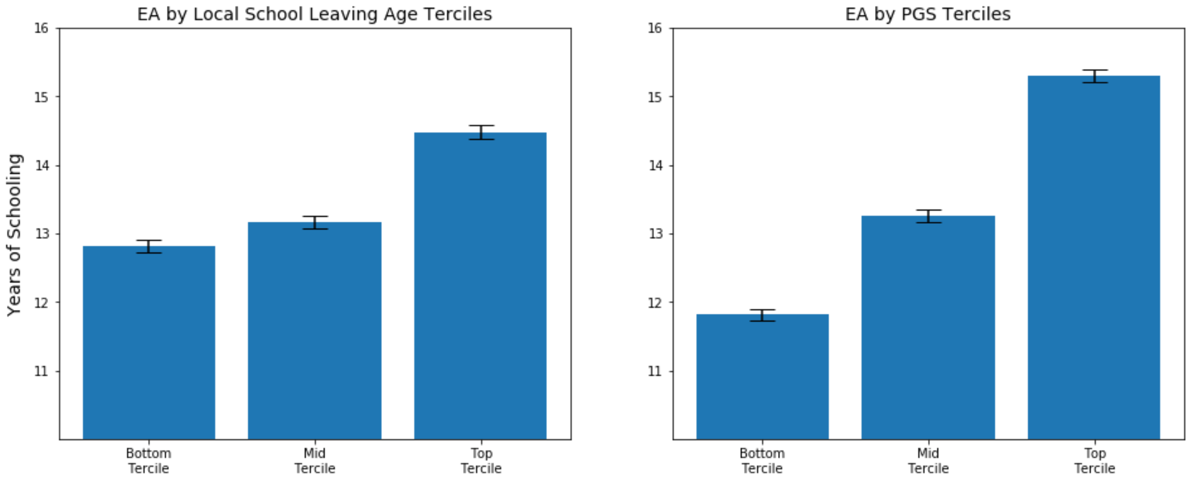
Figure 1 shows the mean educational attainment of the UK Biobank participants, divided into terciles of the distribution of mean local school leaving age in their neighborhood of birth (left panel) and the terciles of the EA PGI distribution (right panel). Participants born in neighborhoods in the top educational attainment terciles have on average 1.7 years more education compared to those in the bottom tercile. Inequality reflected by genetic differences are even larger, with those in the top tercile of the PGI distribution having on average 3.5 years more education than those in the bottom tercile.

Figure 2 shows the results for imputed log hourly wages. Participants in the top local schooling terciles have 1.07 pounds per hour higher wages than those in the bottom, and those born with a genetic endowment in the top tercile have 2.49 pounds per hour higher wages than those in the bottom.

Finally, Figure 3 shows the results for BMI. Participants in the top local schooling tercile have a mean BMI that is 0.53 lower than those in the bottom. The difference between the top and bottom BMI PGI terciles is 3.58 BMI points. For a person that is 180 cm tall, a difference of 3.58 BMI points would be equivalent to 11.6 kilograms.

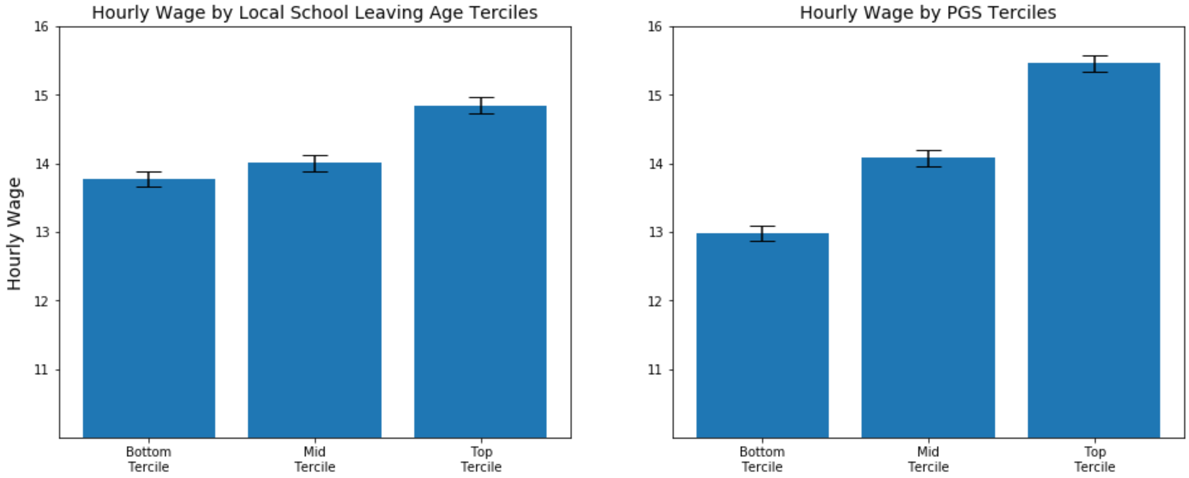
SI Figure 1 shows the mean educational attainment, hourly wage and BMI by terciles of the infant mortality rate distribution. Those results show a similar pattern where persons born in the top tercile have more favorable outcomes than those in the bottom.

Figure 1. Mean of educational attainment for different terciles of the local school leaving age and PGI distribution



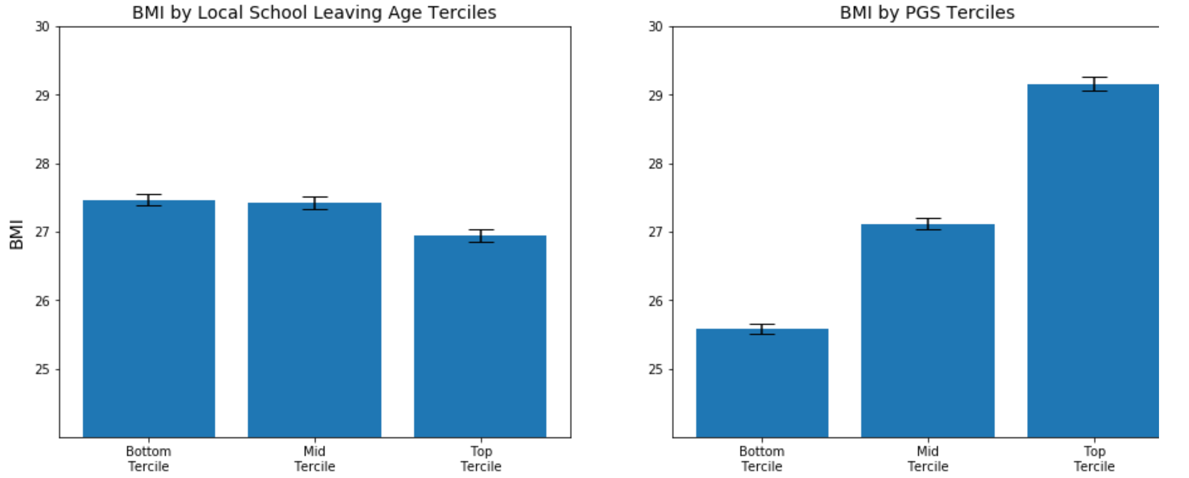
This figure shows the mean of educational attainment (EA), measured in years of schooling, by the different terciles of the average local school leaving age distribution (left panel) and the PGI for EA (right panel).

Figure 2. Mean hourly wage for different terciles of the local school leaving age and PGI distribution



This figure shows the mean of imputed occupational wages, measured in pounds per hour, by the different terciles of the average local school leaving age distribution (left panel) and the PGI for occupational wages (right panel).

Figure 3. Mean BMI for different terciles of the local school leaving age and PGI distribution

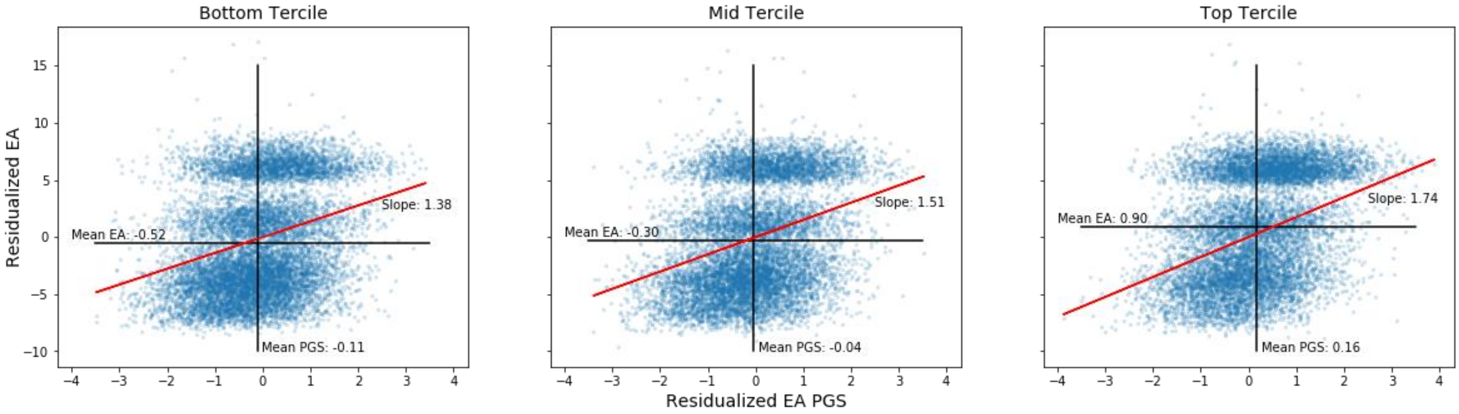


This figure shows the mean body mass index (BMI), by the different terciles of the average local school leaving age distribution (left panel) and the PGI for BMI (right panel).

We illustrate the regression results from model (1) for EA in Figure 4. The panels show the scatterplots of EA and the EA PGI for the bottom, middle, and top terciles of the local school leaving age distribution, after residualizing both axis on control variables. The mean of both EA and the EA PGI vary across environments, an ANOVA mean comparison test shows that they significantly differ from each other ($p \leq 0.0001$). The difference in means show that being born in a district in the top tercile of the local school leaving age distribution is associated with approximately 1.1 years more education compared to being born in the bottom tercile. There is also difference in the effect of the EA PGI on EA between the three local school leaving age terciles, as indicated by a difference in slopes of the regression line. The slope indicates that that a 1 standard deviation (SD) increase in the PGI is associated with an increase in education of approximately 1.4 years in the bottom tercile and

1.51 or 1.74 for the middle or top tercile. The effect of the PGI is stronger for individuals from districts with a higher average school leaving age. Thus, the interaction between the PGI and the district of birth exacerbates the inequalities from each of the two sources of luck. Finally, the mean PGI also varies by the district terciles and its mean is 0.26 higher in the in the top tercile compared to the bottom ($p \leq 0.0001$). Thus, the social and genetic sources of luck that we investigated are positively correlated with each other, which further increases inequalities that arise from them. The results of this regression are also reported in column (1) of table 1. Here, the effects of the EA PGI, the tercile dummies, as well as their interactions are all statistically significant with p-values below 0.05.

Figure 4. Educational attainment by terciles of the local school leaving age

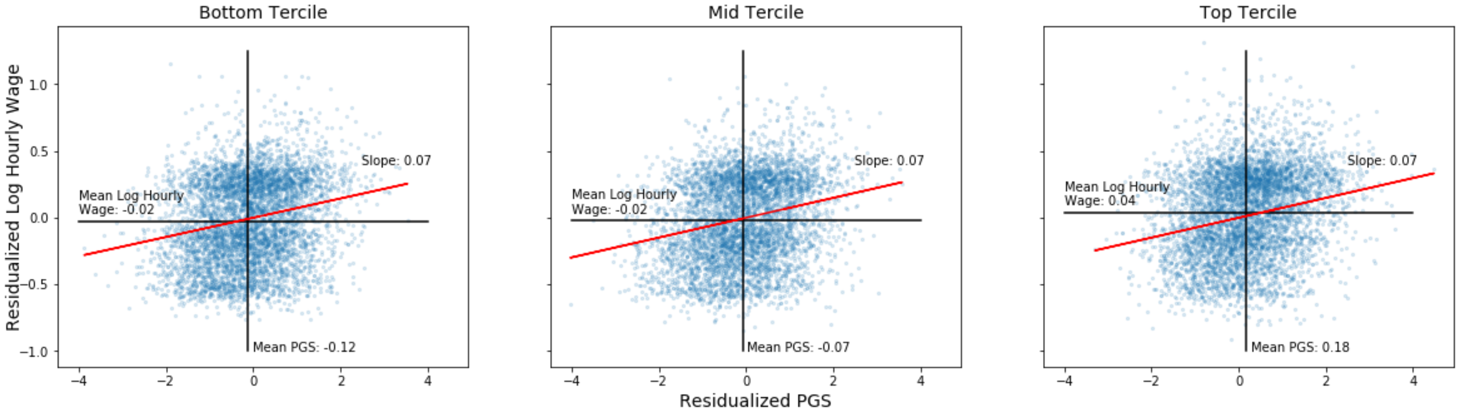


This figure shows the effect of the EA PGI on educational attainment for different terciles of the local school leaving age distribution. We residualized educational attainment and the EA PGI by regressing them on year of birth, year of birth squared, year of birth cubed, gender, gender interacted with the year of birth variables, twenty principal components and genotyping batch.

Column (2) of table 1 reports regression results including family fixed effects. The family fixed effects absorb a substantial part of the signal from the other variables. The district terciles are designed to capture early childhood environmental effects, but the results show that they are not predictive beyond the family environment. It should be noted that the family-fixed effects may capture most of the variation in the terciles and the variation in the local average school leaving age between siblings is typically small, which decreases power. The coefficient for the PGI remains statistically significant ($p \leq 0.001$), but with a lower coefficient. This is in line with previous findings, and may be attributed to genetic nurture effects (Kong et al., 2018; Lee et al., 2018), which are indirect effects from parental genotypes on the offspring through the environment they provide. Parental genotypes are correlated with the genotypes of their offspring, and may also be correlated to environment they provide for their offspring. This induces an unobserved variable bias in the estimated effect of the PGI when there are no controls for parental genotype or family fixed effects. Another possibility is that the PGI also captures some population stratification (Hamer & Sirota, 2000), which is a term that describes the systematic differences in allele frequencies between subpopulations. This could cause an inflation of the coefficient for the PGI the coefficient before controlling for family-fixed effects, if there are environmental differences between the subpopulations that are correlated with the PGI and the outcome, even though twenty principal components of the genetic data were added as control variables (Price et al., 2006). While these two potential sources of the decrease in PGI are not indicative of luck due to direct genetic effects, both still refer to luck due to exogenously given endowments that our outside of one's control (i.e. parental environment and population effects). Nevertheless, our results indicate that direct genetic luck still plays an important role in educational attainment, even when family fixed effects are controlled for: A one standard deviation increase in the PGI implies an increase of 0.8 years of education.

Figure 5 shows the results for imputed log hourly wage and its respective PGI. The regression coefficients are reported in column (3) of table 1. When comparing the bottom to the middle tercile, we see that being born into a middle tercile SES district does not affect hourly wages compared to the bottom tercile. Being born in the top tercile does have an impact on hourly wages and the coefficient in column (3) of table 1 shows that it is associated with 5.0% higher wages. The coefficient for the PGI in column (3) is significant ($p \leq 0.001$) and indicates an increase in wages of approximately 7.2% per SD of the PGI, which is in line with previous findings (Kweon et al., 2020). The relationship between the PGI and mean log hourly wages is identical across terciles, which can be seen from the identical slope of the regression line across terciles in figure 5 and interaction coefficients in column (3). Again, the mean PGI is higher in the highest tercile ($p \leq 0.0001$), indicating a positive correlation between genetic and social luck.

Figure 5. Log hourly wage by terciles of the local school leaving age



This figure shows the effect of the PGI for log hourly wage on imputed log hourly wage for different terciles of the local school leaving age distribution. We residualized log hourly wages and the income PGI by regressing them on year of birth, year of birth squared, year of birth cubed, gender, gender interacted with the year of birth variables, twenty principal components and genotyping batch.

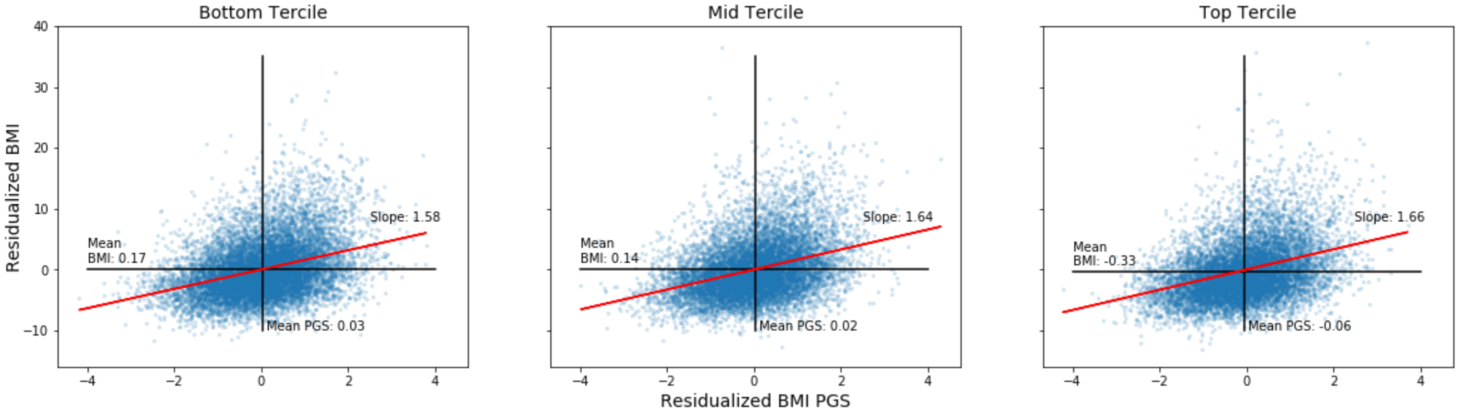
When adding family-fixed effects to the model in column (4), the coefficient of the PGI decreases compared to the model without family-fixed effects, but remains an important predictor of hourly wages. A one standard deviation increase in the PGI is associated with a 5.1% increase in hourly wages. Similar to the effects for EA, the district terciles are not predictive when controlling for family-fixed effects.

The regression results for BMI are visualized in a similar fashion in figure 6. The top tercile of the local school leaving age distribution exhibits a substantially lower average BMI of 0.41 points ($p \leq 0.001$). Furthermore, the BMI PGI is associated with a 1.5 point increase in BMI per standard deviation of the PGI ($p \leq 0.001$). Similar to the results for hourly wages, the relationship between the PGI and BMI does not vary by district tercile.

Comparing the results of column (5) to column (6) in table 1, the coefficient of the BMI PGI barely changes when family fixed-effects are controlled for, which indicates that genetic nurture is less important for BMI than for socio-economic outcomes. This is consistent with the findings reported by Kong et al. (2018) However, controlling for family-fixed effects again absorbs the effects of the local school leaving age on BMI.

We obtained qualitatively similar results when we used the local infant mortality rate terciles as proxies of early childhood environment (SI Table 3). One notable difference is that we observe an interaction effect between the BMI PGI and the local infant mortality rate on BMI in adulthood: The BMI PGI is more strongly associated with BMI in districts with low infant mortality rate. This interaction effect remains even after family-fixed effects are controlled for, indicating that the infant mortality rate may capture health-relevant environmental effects that are not captured by the local average schooling leaving age.

Figure 6. BMI by terciles of the local school leaving age



This figure shows the effect of the PGI for BMI on BMI for different terciles of the local school leaving age distribution. We residualized BMI and the BMI PGI by regressing them on year of birth, year of birth squared, year of birth cubed, gender, gender interacted with the year of birth variables, twenty principal components and genotyping batch.

Finally, the results of our random effects models (equation 2) are shown in Table 2. The estimates for EA, log hourly wage and BMI are shown in columns (1), (2), and (3), respectively. The results are very similar to the models with family fixed effects in Table 1. We see that genetic luck, measured by the respective PGI, remains an important factor even after controlling for non-genetic confounds such as population stratification and environmentally mediated indirect genetic effects from parents on their children. Thus, we find that genetic luck in the form of random genetic differences between siblings is an important factor that contributes to inequalities in socio-economic outcomes and BMI. Specifically, a one standard deviation increase in the PGI for EA is associated with a 0.8 year increase in EA. Similarly, a one standard deviation increase in the PGI for hourly wage is associated with a 4.7% wage increase. Finally, a one standard deviation increase in the PGI for BMI is associated with a 1.6 point increase in BMI. Similar to the family-fixed effects models in table 1, the local school leaving age and the interaction terms lose their predictiveness when all the controls for between-family differences are added.

The overall variance explained in outcomes by our models in Table 2 can be interpreted as a lower bound of the effects of luck because all covariates measure exogenously given endowments that are out of the control of the individual. This includes the outcomes of the social lottery (i.e. the identity of one's parents, the family one is born into, the neighborhood the family lives in) as well the outcomes of the genetic lottery (i.e. one's biological sex and values of the polygenic indices).

Although one does not have any control over their year of birth, it could be argued that year of birth effects should not be counted as luck, as our outcomes may partly be determined by the process of aging. For instance, older employees may have more experience, which may increase their wages. Furthermore, biological processes in our body change due to aging which may affect our BMI. As everyone will go through the process of aging in their life, it could be argued that this should not be attributed to luck. However, it should be noted that the UK Biobank participants are past the typical schooling age, as the participants were between 40 and 69 when they entered the study. Thus, if birth year has an effect on educational attainment, it could very well be due to the luck of the social circumstances one is born in, as different schooling regulations were in place depending on the exact age of the participants.

To re-evaluate the amount of variance explained in our outcomes that can be attributed to sources of luck, we re-estimate our models from Table 2 by first removing all birthyear effects. Here, we first regress our outcomes and all covariates on birth year, birth year squared and birth year cubed and take the residuals. In these models approximately 14 percent of educational attainment can be attributed to luck, 17 percent of occupational wages and 13 percent of BMI. When comparing these numbers to the overall R^2 measures in Table 2, we see that the change in the share of variation that can be attributed to luck does not change much. The largest change is for educational attainment, where the share of luck drops by approximately 2 percent.

SI table 4 shows the results of the random effects models using the local infant mortality rate as an early life exposure. The results are very similar to those reported in table 2. For each of the outcomes, the PGI has a similar effect size to those reported in table 2, and the local infant mortality rate and interaction terms are not predictive of the outcomes.

Table 1. Regression of the outcomes on PGI, local school leaving age terciles and interactions

	EA		Log Hourly Wage		BMI	
	(1)	(2)	(3)	(4)	(5)	(6)
Family Fixed Effects	No	Yes	No	Yes	No	Yes
PGI	1.384	0.799	0.072	0.051	1.584	1.561
S.E.	0.043	0.068	0.004	0.008	0.040	0.066
p-value	0.000	0.000	0.000	0.000	0.000	0.000
Middle Tercile	0.190	-0.072	0.006	-0.018	-0.045	0.173
S.E.	0.062	0.139	0.006	0.016	0.060	0.154
p-value	0.002	0.606	0.375	0.272	0.453	0.261
Top Tercile	1.123	-0.139	0.050	-0.003	-0.412	0.074
S.E.	0.065	0.144	0.007	0.016	0.063	0.150
p-value	0.000	0.334	0.000	0.838	0.000	0.619
PGI x Middle Tercile	0.124	0.033	0.002	-0.009	0.057	0.126
S.E.	0.061	0.092	0.006	0.011	0.059	0.095
p-value	0.043	0.719	0.713	0.409	0.328	0.185
PGI x Top Tercile	0.360	0.060	0.003	-0.005	0.077	0.033
S.E.	0.061	0.092	0.006	0.010	0.059	0.091
p-value	0.000	0.511	0.573	0.654	0.191	0.717
R^2 (Overall)	0.161	0.107	0.175	0.144	0.137	0.126
R^2 (Between)		0.136		0.145		0.146
R^2 (Within)		0.039		0.143		0.092
N	32474	32474	16175	16175	32942	32942
Sibling Groups		15787		7894		16013

This table shows Ordinary Least Squares (OLS) regression results for regressing each of the outcomes (Educational Attainment (EA), Imputed log hourly wages and Body Mass Index (BMI)), on their respective polygenic indices (PGI), local school leaving age terciles and interactions. Columns 1 and 2 show results for EA, columns 3 and 4 for log hourly wage, and columns 5 and 6 for BMI. Columns 2, 4 and 6 include family fixed effects. All regressions included the following control variables: age, age squared, age cubed, gender, gender interacted with age variables, twenty principal components of the genetic data and dummies for genotyping batches. In the family fixed effects models some control variables had to be dropped due to multi-collinearity.

Table 2. Random Effects Models

	EA	Log Hourly Wage	BMI
	(1)	(2)	(3)
PGI	0.823	0.047	1.612
S.E.	0.043	0.005	0.042
p-value	0.000	0.000	0.000
Local School Leaving Age	-0.031	0.010	-0.065
S.E.	0.085	0.009	0.087
p-value	0.712	0.260	0.454
PGI x Local School Leaving Age	-0.133	0.012	-0.018
S.E.	0.112	0.029	0.112
p-value	0.238	0.684	0.872
R^2 (Overall)	0.163	0.168	0.132
R^2 (Between)	0.216	0.188	0.156
R^2 (Within)	0.036	0.140	0.090
N	32474	16175	32942
Sibling Groups	15787	7894	16013

This table shows the results of the random effects models based on a Mundlak formulation. Column (1) shows results for educational attainment (EA), column (2) for imputed log hourly wages, column (3) for body mass index (BMI). The outcomes were regressed on the PGI, Local School Leaving age, their interaction and control variables (gender, year of birth and year of birth squared). For each variable within-family means were added to control for between family variation. See equation 2 for the full model.

5. Discussion

We investigated the effects of inequalities in opportunity that are due to social and genetic luck on educational attainment, occupational wages and BMI. We tested potential interaction effects between genes and environment in a novel within-family study design that uses the random genetic differences between siblings to break the link between family-environments and genes. This approach allowed us to obtain estimates of gene-environment interactions that do not suffer from endogeneity bias, which is a common concern in gene-environment studies (Harden & Koellinger, 2020; Schmitz & Conley, 2017a).

Our results illustrate that both social and genetic luck contribute towards inequalities in socio-economic status and BMI. Our estimates suggest that at least 13 to 17 percent of the inequalities in education, wages and BMI in the UK are due to inequalities in opportunity that arise from the outcomes of the social and the genetic lottery. This estimate is likely to be strongly attenuated by measure error both in the polygenic indices and the proxies of childhood environment that we had available. Thus, the true influence of social and genetic luck on inequalities in the UK is likely to be substantially higher in reality. Future investigations on this would benefit from more precise polygenic indices as well as better measures of relevant environments during childhood.

Our results also showed that social and genetic luck are correlated, which exacerbates their influence on disparities on socioeconomic and health outcomes. This type of Matthew effect had previously been identified in large-scale, genetically informed study designs. For example, children who grew up in high-SES households also tended to have higher polygenic indices for educational attainment in Belsky et al. (2018). In Abdellaoui et al. (2019), polygenic index values for educational attainment where on average lower in regions of the UK that had overall lower SES (e.g. former coal mining areas). Furthermore, the indirect genetic effects for educational attainment reported by Kong et al. (2018) are another example for how tightly intertwined genetic and environmental factors are that contribute towards SES.

Our results further emphasize the importance of both social and genetic luck as drivers of inequalities in socio-economic status and BMI. In particular, we found that genetic luck is a strong predictor for our outcomes in all our model specifications, including those that rely on the random genetic differences between siblings for identification (e.g. models in which genetic effects have a causal interpretation). In contrast, we find that the early childhood environmental exposures lose their predictiveness when we control for family-fixed effects. Similarly, we find some evidence for gene-by-environment interactions, but not when controlling for family-specific effects.

However, this does not imply that social luck is less important than genetic luck. Rather, the environmental exposures we studied are based on noisy neighbourhood proxies that are unlikely to capture all facets of the environment that are relevant. Moreover, siblings are typically born in similar socio-economic environments and are on average 50% genetically identical. This implies that the differences between siblings tend to be smaller than differences between unrelated individuals, which decreases statistical power to detect true effects in study designs such as ours that use the random differences between siblings for identification. Even larger samples of genotyped siblings would be desirable to identify relevant environment and gene-by-environment interactions in such study designs. Thus, our study illustrates some of the challenges for identifying robust, non-endogenous gene-environment interactions.

The relative importance of social and genetic luck that we studied here has policy relevance because the extent to which people are willing to tolerate or endorse inequality partially depends on whether they perceive that disparity originates from differences in effort and choice (e.g., working hard) or from differences in circumstances that are outside of one's control (e.g., luck in the social or genetic lotteries). The existing empirical evidence suggests that inequality that can ultimately be traced back to luck may be perceived as unfair and people may favor redistributive policies more strongly if inequality is the result of luck rather than agency (Alesina & La Ferrara, 2005; Alesina et al., 2018; Almås et al., 2010; Cappelen et al., 2013; Clark & D'Ambrosio, 2015; Gromet et al., 2015). Furthermore, policies that aim at providing broad access to education and health care are desirable if policies aim at providing people with equal opportunities. However, more equal opportunities do not necessarily translate into equalities in outcomes. For example, previous studies have indicated that schooling reforms can reduce disparities in education and health that are rooted in genetic effects, but these reforms may not decrease inequalities in wages (Barcellos et al., 2018, 2020). Thus, it is important for science and policy to better understand the extent to which genetic and social luck contribute to inequality, the mechanisms that are at work, and whether and how the consequences of exogenously given endowments can be altered.

Author Contributions

Conceived and designed the study: CAPB, HK, PDK. Prepared data: CAPB, HK. Analysed the data: CAPB. Wrote and edited the manuscript: CAPB, PDK, HK.

Acknowledgments

This research was conducted using the UK Biobank Resource under applications 11425. We thank all UKB cohort participants for making this study possible. This study was financially supported through a European Research Council Consolidator Grant to PDK (647648 EdGe). Support for this research was provided by the Office of the Vice Chancellor for Research and Graduate Education at the University of Wisconsin –Madison with funding from the Wisconsin Alumni Research Foundation. These analyses were made possible by the generous public sharing of summary statistics from published GWAS from the Social Science Genetics Association Consortium (SSGAC) and the Genetic Investigation of Anthropometrics Traits Consortium (GIANT). We would like to thank the many studies that made these consortia possible, the researchers involved, and the participants in those studies, without whom this effort would not be possible.

Data reporting

Access to the UK Biobank resource can be requested https://bbams.ndph.ox.ac.uk/ams/. We will share the polygenic indices we created and the district information we obtained via the standard data sharing mechanism of the UK Biobank.

References

1000 Genomes Project Consortium, T. 1000 G. P., Auton, A., Brooks, L. D., Durbin, R. M., Garrison, E. P., Kang, H. M., … Abecasis, G. R. (2015). A global reference for human genetic variation. *Nature*, *526*(7571), 68–74. https://doi.org/10.1038/nature15393

Abdellaoui, A., Hugh-Jones, D., Kemper, K. E., Holtz, Y., Nivard, M. G., Veul, L., … Visscher, P. M. (2019). Genetic correlates of social stratification in Great Britain. *Nature Human Behaviour*, *3*, 1332–1342. https://doi.org/https://doi.org/10.1038/s41562-019-0757-5

Adler, N. E., & Rehkopf, D. H. (2008). U.S. disparities in health: Descriptions, causes, and mechanisms. *Annual Review of Public Health*, *29*(1), 235–252. https://doi.org/10.1146/annurev.publhealth.29.020907.090852

Alesina, A., & La Ferrara, E. (2005). Preferences for redistribution in the land of opportunities. *Journal of Public Economics*, *89*(5), 897–931. https://doi.org/https://doi.org/10.1016/j.jpubeco.2004.05.009

Alesina, A., Stantcheva, S., & Teso, E. (2018). Intergenerational Mobility and Preferences for Redistribution. *American Economic Review*, *108*(2), 521–554. https://doi.org/10.1257/aer.20162015

Almås, I., Cappelen, A. W., Sørensen, E. Ø., & Tungodden, B. (2010). Fairness and the development of inequality acceptance. *Science (New York, N.Y.)*, *328*(5982), 1176–1178. https://doi.org/10.1126/science.1187300

Barcellos, S. H., Carvalho, L. S., & Turley, P. (2018). Education can reduce health differences related to genetic risk of obesity. *Proceedings of the National Academy of Sciences*, *115*(42), E9765 LP-E9772. https://doi.org/10.1073/pnas.1802909115

Barcellos, S. H., Carvalho, L. S., & Turley, P. (2020). *Is Education the Great Equalizer?* Retrieved from http://www2.nber.org/conferences/2020/SI subs/Equalizer_NBER1.pdf

Becker, J., Burik, C. A. P., Goldman, G., Wang, N., Jayashankar, H., Bennett, M., … Okbay, A. (2021). Resource Profile and User Guide of the Polygenic Index Repository. *Nature Human Behaviour*, *Forthcomin*.

Belsky, D. W., Domingue, B. W., Wedow, R., Arseneault, L., Boardman, J. D., Caspi, A., … Harris, K. M. (2018). Genetic analysis of social-class mobility in five longitudinal studies. *Proc Natl Acad Sci U S A*, *115*(31), E7275–E7284. https://doi.org/10.1073/pnas.1801238115

Benjamin, D. J., Cesarini, D., van der Loos, M. J. H. M., Dawes, C. T., Koellinger, P. D., Magnusson, P. K. E., … Visscher, P. M. (2012). The genetic architecture of economic and political preferences. *Proceedings of the National Academy of Sciences*, *109*(21), 8026–8031. https://doi.org/10.1073/pnas.1120666109

Boyce, C. J., Brown, G. D. A., & Moore, S. C. (2010). Money and happiness: Rank of income, not income, affects life satisfaction. *Psychological Science*. Boyce, Christopher J.: Department of Psychology, University of Warwick, Gibbet Hill Rd., Coventry, United Kingdom, CV4 7AL, c.j.boyce@warwick.ac.uk: Sage Publications. https://doi.org/10.1177/0956797610362671

Cappelen, A. W., Konow, J., Sørensen, E. Ø., & Tungodden, B. (2013). Just Luck: An Experimental Study of Risk-Taking and Fairness. *American Economic Review*, *103*(4), 1398–1413. https://doi.org/10.1257/aer.103.4.1398

Chetty, R., & Hendren, N. (2018a). The Impacts of Neighborhoods on Intergenerational Mobility I: Childhood Exposure Effects. *Quarterly Journal of Economics*, *113*(3). Retrieved from https://academic.oup.com/qje/article-abstract/133/3/1107/4850660

Chetty, R., & Hendren, N. (2018b). The Impacts of Neighborhoods on Intergenerational Mobility II: County-Level Estimates*. *The Quarterly Journal of Economics*, *133*(3), 1163–1228. https://doi.org/10.1093/qje/qjy006

Clark, A., & D'Ambrosio, C. (2015). Attitudes to Income Inequality: Experimental and Survey Evidence. *Handbook of Income Distribution*, *2*. https://doi.org/10.1016/B978-0-444-59428-0.00014-X

Daetwyler, H. D., Villanueva, B., & Woolliams, J. a. (2008). Accuracy of predicting the genetic risk of disease using a genome-wide approach. *PLoS One*, *3*(10), e3395. https://doi.org/10.1371/journal.pone.0003395

de Vlaming, R., Okbay, A., Rietveld, C. A., Johannesson, M., Magnusson, P. K. E., Uitterlinden, A. G., … Koellinger, P. D. (2017). Meta-GWAS Accuracy and Power (MetaGAP) Calculator Shows that Hiding Heritability Is Partially Due to Imperfect Genetic Correlations across Studies. *PLoS Genetics*, *13*(1), e1006495. https://doi.org/10.1371/journal.pgen.1006495

DiPrete, T. A., Burik, C. A. P., & Koellinger, P. D. (2018). Genetic instrumental variable regression: Explaining socioeconomic and health outcomes in nonexperimental data. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(22). https://doi.org/10.1073/pnas.1707388115

Dudbridge, F. (2013). Power and Predictive Accuracy of Polygenic Risk Scores. *PLoS Genetics*, *9*(3). https://doi.org/10.1371/journal.pgen.1003348

Frijters, P., Haisken-DeNew, J. P., & Shields, M. A. (2004). Money Does Matter! Evidence from Increasing Real Income and Life Satisfaction in East Germany Following Reunification. *American Economic Review*, *94*(3), 730–740. https://doi.org/10.1257/0002828041464551

Fry, A., Littlejohns, T. J., Sudlow, C., Doherty, N., Adamska, L., Sprosen, T., … Allen, N. E. (2017). Comparison of Sociodemographic and Health-Related Characteristics of UK Biobank Participants With Those of the General Population. *American Journal of Epidemiology*, *186*(9), 1026–1034. https://doi.org/10.1093/aje/kwx246

Giesselmann, M., & Schmidt-Catran, A. W. (2018). Getting the Within Estimator of Cross-Level Interactions in Multilevel Models with Pooled Cross-Sections: Why Country Dummies (Sometimes) Do Not Do the Job. *Sociological Methodology*, *49*(1), 190–219. https://doi.org/10.1177/0081175018809150

Giesselmann, M., & Schmidt-Catran, A. W. (2020). Interactions in Fixed Effects Regression Models. *Sociological Methods & Research*, 0049124120914934. https://doi.org/10.1177/0049124120914934

Gromet, D. M., Hartson, K. A., & Sherman, D. K. (2015). The politics of luck: Political ideology and the perceived relationship between luck and success. *Journal of Experimental Social Psychology*, *59*, 40–46. https://doi.org/https://doi.org/10.1016/j.jesp.2015.03.002

Hamer, D. H., & Sirota, L. (2000). Beware the chopsticks gene. *Molecular Psychiatry*, *5*(1), 11–13. https://doi.org/10.1038/sj.mp.4000662

Harden, K. P., & Koellinger, P. D. (2020). Using genetics for social science. *Nature Human Behaviour*.

Hill, W. D., Davies, N. M., Ritchie, S. J., Skene, N. G., Bryois, J., Bell, S., … Deary, I. J. (2019). Genome-wide analysis identifies molecular systems and 149 genetic loci associated with income. *Nature Communications*, *10*(1), 5741. https://doi.org/10.1038/s41467-019-13585-5

Kong, A., Thorleifsson, G., Frigge, M. L., Vilhjalmsson, B. J., Young, A. I., Thorgeirsson, T. E., … Stefansson, K. (2018). The nature of nurture: Effects of parental genotypes. *Science*, *359*(6374), 424–428. https://doi.org/10.1126/science.aan6877

Kweon, H., Burik, C. A. P., Karlsson Linnér, R., de Vlaming, R., Okbay, A., Martschenko, D., … Koellinger, P. D. (2020). *Genetic Fortune: Winning or Losing Education, Income, and Health* (Tinbergen Institute Discussion Papers No. 20- 053/V). Amsterdam.

Lee, J. J., Wedow, R., Okbay, A., Kong, E., Maghzian, O., Zacher, M., … others. (2018). Gene discovery and polygenic prediction from a 1.1-million-person GWAS of educational attainment. *Nature Genetics*, *50*(8), 1112.

Locke, A. E., Kahali, B., Berndt, S. I., Justice, A. E., Pers, T. H., Day, F. R., … Speliotes, E. K. (2015). Genetic studies of body mass index yield new insights for obesity biology. *Nature*, *518*(7538), 197–206. https://doi.org/10.1038/nature14177

Lu, Y., Day, F. R., Gustafsson, S., Buchkovich, M. L., Na, J., Bataille, V., … Loos, R. J. F. (2016). New loci for body fat percentage reveal link between adiposity and cardiometabolic disease risk. *Nature Communications*, *7*, 10495. https://doi.org/10.1038/ncomms10495

Mokdad, A. H., Ford, E. S., Bowman, B. A., Dietz, W. H., Vinicor, F., Bales, V. S., & Marks, J. S. (2003). Prevalence of obesity, diabetes, and obesity-related health risk factors, 2001. *JAMA*, *289*(1), 76–79. https://doi.org/10.1001/jama.289.1.76

Mundlak, Y. (1978). On the pooling of time series and cross section data. *Econometrica*, *46*(1), 69. https://doi.org/10.2307/1913646

Polderman, T. J. C., Benyamin, B., de Leeuw, C. A., Sullivan, P. F., van Bochoven, A., Visscher, P. M., & Posthuma, D. (2015). Meta-analysis of the heritability of human traits based on fifty years of twin studies. *Nature Genetics*, *advance on*. https://doi.org/10.1038/ng.3285

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., & Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, *38*(8), 904–909. https://doi.org/10.1038/ng1847

Rosenzweig, M. R., & Wolpin, K. I. (2000). Natural "Natural Experiments" in Economics. *Journal of Economic Literature*, *38*(4), 827–874.

Schmitz, L. L., & Conley, D. (2017a). Modeling gene-environment interactions with quasi-natural experiments. *Journal of Personality*, *85*(1), 10–21.

Schmitz, L. L., & Conley, D. (2017b). The effect of Vietnam-era conscription and genetic potential for educational attainment on schooling outcomes. *Economics of Education Review*, *61*, 85–97. https://doi.org/https://doi.org/10.1016/j.econedurev.2017.10.001

Southall, H. (2011). Rebuilding the Great Britain Historical GIS, Part 1: Building an Indefinitely Scalable Statistical Database. *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, *44*(3), 149–159. https://doi.org/10.1080/01615440.2011.589774

Stringhini, S., Carmeli, C., Jokela, M., Avendaño, M., Muennig, P., Guida, F., … Zins, M. (2017). Socioeconomic status and the 25 × 25 risk factors as determinants of premature mortality: a multicohort study and meta-analysis of 1·7 million men and women. *The Lancet*, *389*(10075), 1229–1237. https://doi.org/10.1016/S0140-6736(16)32380-7

Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., … Collins, R. (2015). UK Biobank: An open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Medicine*, *12*(3), e1001779. https://doi.org/10.1371/journal.pmed.1001779

Taubman, P. (1976). The determinants of earnings: Genetics, family, and other environments: A study of white male twins. *The American Economic Review*, *66*(5), 858–870. Retrieved from http://www.jstor.org/stable/1827497

Turley, P., Walters, R. K., Maghzian, O., Okbay, A., Lee, J. J., Fontana, M. A., … Benjamin, D. J. (2018). Multi-trait analysis of genome-wide association summary statistics using MTAG. *Nature Genetics*, *50*(2), 229–237. https://doi.org/10.1038/s41588-017-0009-4

Vilhjálmsson, B. J., Jian Yang, H. K. F., Alexander Gusev, S. L., Stephan Ripke, G. G., Po-Ru Loh, Gaurav Bhatia, R. Do, Tristan Hayeck, H.-H. W., … Price, A. L. (2015). Modeling Linkage Disequilibrium Increases Accuracy of Polygenic Risk Scores. *The American Journal of Human Genetics*, *97*(4), 576–592. https://doi.org/10.1016/j.ajhg.2015.09.001

Weightman, A. L., Morgan, H. E., Shepherd, M. A., Kitcher, H., Roberts, C., & Dunstan, F. D. (2012). Social inequality and infant health in the UK: systematic review and meta-analyses. *BMJ Open*, *2*(3), e000964. https://doi.org/10.1136/bmjopen-2012-000964

Wilkinson, R. G., & Marmot, M. (2003). *Social Determinants of Health: The Solid Facts*. World Health Organization.

SUPPORTING INFORMATION

Disparities in socio-economic status and BMI in the UK are partly due to genetic and environmental luck

Casper A.P. Burik[1], Hyeokmoon Kweon[1] and Philipp D. Koellinger*[2,1]

1. Department of Economics, School of Economics and Business, Vrije Universiteit Amsterdam
2. La Follette School of Public Affairs, University of Wisconsin-Madison
* Corresponding author: koellinger@wisc.edu

1. Table of Contents

2. Measures of early-childhood environment

As early-childhood environmental exposures, we derived the local average school leaving age and the infant mortality rate at the district level by exploiting the birth locations provided by the UKB. We obtained the historical local-level data from Vision of Britain (www.visionofbritain.org.uk), which covers the period from the early 20th century to the 1970s. Using boundary data for local government districts as of 1931, 1951, 1961, and 1971, we first coded the birth locations in terms of local government district. Based on this information, we constructed childhood local environment measures by matching the birth places to the local-level data.

We derived the local average school leaving age as of 1961 by using district-level data provided as fractions of pupils in the district who left school at the age of under-15, 15, 16, 17 to 19, and above-20. To these fractions, we multiplied the values of 10, 15, 16, 18, and 20, respectively, to compute the average school leaving age of the district. This data was only available for 1961.[1] We used the boundary data for local government districts as of 1961 to match the local average school leaving age to each participant.

The local infant mortality rates were available at the district level annually. To reduce the noise in the data, we smoothed the infant mortality rate time series for each district by using the Hodrick-Prescott filter with the smoothing parameter of 100 (Hodrick & Prescott, 1997). We also dropped observations if the number of births in the district was fewer than 50 in that year. The boundary data was only available for 1931, 1951, 1961, and 1971 while the local infant mortality data was available annually. Therefore, we used the boundary data from the year nearest to the birth year for each participant.

3. Polygenic indices

We constructed polygenic indices (PGI) using the results of the largest GWAS that are currently publicly available for educational attainment (Lee et al., 2018), occupational wages (Kweon et al., 2020), and BMI (Locke et al., 2015). We further improved the accuracy of these PGI with MTAG (Turley et al., 2018), which is a multivariate statistical method that increases the statistical power of GWAS by including GWAS summary statistics from genetically correlated phenotypes.

MTAG analyses included GWAS summary statistics of phenotypes that pass the following criteria:

1. The phenotype belongs to the same scientific domain as the outcome variable of interest. This limits the possibilities of spurious associations when covariates are genetically correlated to the outcome.
2. The phenotype has been included in a previously published GWAS, as GWAS for novel phenotypes would go beyond the scope of this paper.
3. Genetic correlation (r(G)) between the phenotypes is at least 0.6 Here we follow the genetic correlation threshold of Becker et al. (2020) Where the authors conduct many MTAG analyses to construct a repository of PGI.

[1] In 1951, this data was only available for men. Therefore, we only used the 1961 data.

4. The heritability of the trait is significantly different from 0. Adding traits with little genetic signal would only add noise to our PGI.
5. The GWAS had a sample size of at least 20,000. So that the phenotype contributes significantly to the predictiveness of the PGI.

SI Table 1 gives an overview of all included GWAS summary statistics that meet these criteria. These studies were found via a systematic literature review and genetic correlations provided by LD Hub (Zheng et al., 2017) and Becker et al. (2020) If the phenotype was available in the UK Biobank, we conducted GWAS on a subsample of the UK Biobank that excluded siblings and their genetic relatives (see section 3). Genetic relatives were identified using relatedness coefficients provided by the UK Biobank. We meta-analysed these results with the publicly available GWAS summary statistics that excluded the UK Biobank. SI Table 1 provides an overview of the GWASs run in the UK Biobank.

SI Table 2 gives an overview of phenotypes that meet the above criteria, but had to be dropped during our preliminary analyses. The table also provides the reason for their dismissal.

To adjust for linkage-disequilibrium, we constructed PGI using LDpred (Vilhjálmsson et al., 2015). The Haplotype Reference Consortium (McCarthy et al., 2016) panel was used as LD reference and we employed the recommended LD window, (number of SNPs divided by 3000) and set the fraction of causal markers to 1. We limited the number of SNPs that we included in the PGI to those that are directly genotyped or are present in the HapMap3 reference panel (International HapMap 3 Consortium et al., 2010). This set of SNPs provides a good coverage of common genetic variants and it tends to yield PGI that perform well empirically (Lee et al., 2018). The number of SNPs included in each PGI is further limited by the fact that MTAG only considers SNPs that are present in all summary statistics. The remaining number of SNPs are 1,209,700; 1,209,700; and 1,188,098 for EA, Occupational wages and BMI respectively.

SI Table 1. Overview of GWAS Summary Statistics

Phenotype	Target	r(G)	N	Source
Hardest Math Class	EA, Occ. Wages	0.81, 0.78	430,439	(Lee et al., 2018)
Cognitive Performance	EA, Occ. Wages	0.63, 0.67	35,298	(Trampush et al., 2017)
Cognitive Performance	EA, Occ. Wages	0.63, 0.67	129,048	UKB Data Field: 20016
Cognitive Performance	EA, Occ. Wages	0.63, 0.67	101,205	UKB Data Field: 20191
Household Income	EA, Occ. Wages	0.74, 0.91	340,935	(Kweon et al., 2020)
Regional Income	EA, Occ. Wages	0.81, 0.83	359,437	(Kweon et al., 2020)
Body fat percentage	BMI	0.84	390,601	UKB Data Field: 23099
Hip Circumference	BMI	0.87	224,459	(Shungin et al., 2015)
Hip Circumference	BMI	0.87	397,156	UKB Data Field: 49
Waist Circumference	BMI	0.90	224,459	(Shungin et al., 2015)
Waist Circumference	BMI	0.90	397,197	UKB Data Field: 48

This table gives an overview of GWAS summary statistics from previous studies used to improve the accuracy of the **PGI**. The first column states the phenotype of the GWAS. The second column indicates for which outcome the summary statistics were used. The third column gives the genetic correlation between the phenotype and target outcome. The genetic correlation was calculated using the meta-analysed results if there were multiple sources for that phenotype. The reported correlation was calculated during our preliminary MTAG analyses. The fourth column gives the size of the GWAS. The fifth column gives a reference to the study where the GWAS was published or the UKB Data-Field.

SI Table 2. Overview of Dismissed GWAS Summary Statistics

Phenotype	Target	Source	Reason for dismissal
College Completion	EA, Occ. Wages	(Rietveld et al., 2013)	A
Body fat percentage	BMI	(Lu et al., 2016)	B
Obesity Class 1	BMI	(Berndt et al., 2013)	A
Obesity Class 2	BMI	(Berndt et al., 2013)	A
Obesity Class 3	BMI	(Berndt et al., 2013)	A
Overweight	BMI	(Berndt et al., 2013)	A
Leptin	BMI	(Kilpeläinen et al., 2016)	C
HOMA-IR	BMI	(Dupuis et al., 2010)	D

This table gives an overview of GWAS summary statistics but were dropped in preliminary analyses. The first column states the phenotype of the GWAS. The second column indicates for which outcome the summary statistics were used. The third column gives a reference to the study where the GWAS was published. The fourth column gives the reason code for the dismissal. Where the codes are as follows: A: the phenotype is a binary measure of another included phenotype and the sample is completely overlapping with it. B: the results are from mixed ancestry. C: The phenotype greatly reduced the number of overlapping SNPs used by MTAG. D: the phenotype had no reported number of samples per SNP.

4. GWAS in the UKB

For the phenotypes indicated with the UKB as the source in Table 1, we conducted GWAS on the UKB participants of European ancestry excluding those in the sibling sample and their close relatives (up to the third degree).

We followed the standard phenotype definitions in the literature except for the income outcomes. We coded household income as the natural log of the midpoint income of each income bracket, where 3/4 times the upper bound and 4/3 times the lower bound were used as the midpoint respectively for the lowest and highest brackets, which are open-ended. Regional income (local average weekly household income in 2011) was derived from home locations coded in Middle-layer Super Output Areas. We obtained the income data from the UK's Office for National Statistics, which was available for England and Wales only.

For the non-income outcomes, the control variables included dummy variables for sex, age, year of observation, and assessment centre, and their interaction with sex dummy as well as genotyping arrays and batches and 40 top genetic principal components. For the income outcomes, we conducted GWAS on male and female samples separately and meta-analysed the male and female results of each measure by relying on the meta-analysis version of MTAG to address possible sex heterogeneity in economic outcomes. In the GWAS of the income outcomes, dummy variables for employment status were additionally included.

Each GWAS was run based on a linear mixed model, estimated with BOLT-LMM (Loh et al., 2015). We then applied standard quality control filters to exclude SNPs that are problematic, which we implemented with EasyQC (Winkler et al., 2014). These filters removed SNPs that had missing or incorrect numerical values for output statistics (a p-value outside of [0,1], for example); duplicate SNPs; imputation accuracy below 0.7; a minor allele frequency lower than 0.1%; an allele other than "A," "C," "G," or "T"; or had an allele frequency that deviates 0.2 or more from the allele frequency in the reference panel (Haplotype Reference Consortium v1.1 (McCarthy et al., 2016)).

5. Measuring educational attainment

We measured educational attainment in years of schooling, using a transformation from the highest achieved diploma to a set number of years such that it retains the rank order of lowest to highest degree as much as possible (see SI Table 3). Because the participants could report more than one qualifications, each reported qualification was converted to years of schooling and the maximum value was retained.

SI Table 3. Transformation Qualification to Years of Schooling

Qualification	**Years of schooling**
College or University degree	20
A levels/AS levels or equivalent	13
O levels/GCSEs or equivalent	10
CSEs or equivalent	10
NVQ or HND or HNC or equivalent	Age when left full-time education – 5
Other professional qualifications e.g.: nursing, teaching	15
None of the above	7

This table shows the conversion for each type of diploma to a set years of schooling

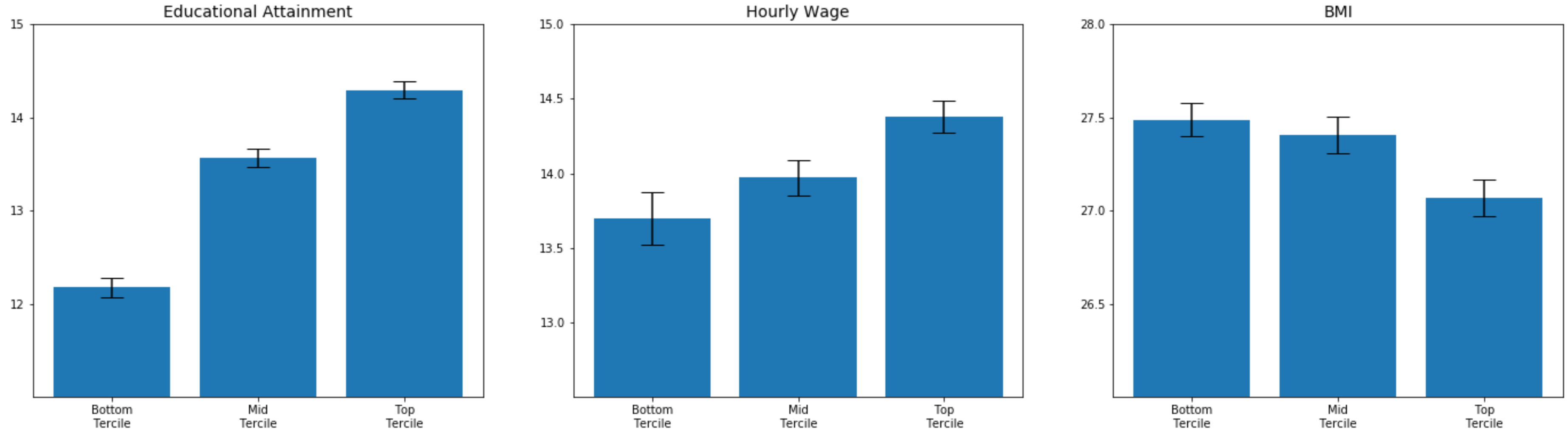
6. Results for infant mortality rate

This section shows the results when using infant mortality rate as early childhood environmental exposure. SI Figure 1 shows the mean of EA, hourly wages and BMI of the UK Biobank participants when divided into terciles of the infant mortality rate distribution. Infant mortality rate was reverse coded such that higher numbers are better to ease the comparison to the results using local school leaving age, as discussed in the main text. The results for EA and BMI are similar to the results using local school leaving age. For hourly wages, we note the differences in sample sizes for the different terciles of the distribution. There are many more missing observations for hourly wages in the bottom tercile of the distribution, indicating attrition in our sample. Thus, the results for hourly wages cannot be interpreted in any meaningful way.

SI Table 3 shows the equivalent results of table 1 in the main text using local infant mortality rate terciles instead of local school leaving age. The results are in line with those of table 1. One notable difference is the interaction effects for BMI. There we do find that the PGI is more strongly associated with BMI in neighbourhoods with low infant mortality rate. The interaction effect remains for the middle tercile, even when controlling for family fixed effects. Again, due to attrition in the sample the results for hourly wages cannot be interpreted in any meaningful way.

SI table 4 shows the results for the random effects models using the local infant mortality rate as an early life exposure. The results are very similar to that of table 2 in the main text.

SI Figure 1. Outcomes by infant mortality rate terciles



This figure shows each of the outcomes plotted by the district infant mortality rate terciles. Infant mortality rate is reverse coded such that higher numbers are good. The left panel shows educational attainment, the middle hourly wages and the right BMI.

SI Table 4. Regression of the outcomes on PGI, infant mortality rate terciles and interactions

	EA		Log Hourly Wage		BMI	
	(1)	(2)	(3)	(4)	(5)	(6)
Family Fixed Effects	No	Yes	No	Yes	No	Yes
PGI	1.483	0789	0.082	0.071	1.461	1.512
S.E.	0.049	0.071	0.008	0.012	0.046	0.068
p-value	0.000	0.000	0.000	0.000	0.000	0.000
Middle Tercile	0.644	0.034	0.016	0.014	-0.082	0.122
S.E.	0.087	0.113	0.011	0.015	0.084	0.115
p-value	0.002	0.003	0.138	0.346	0.330	0.286
Top Tercile	0.706	-0.026	0.025	0.007	-0.504	0.029
S.E.	0.124	0.177	0.014	0.020	0.120	0.181
p-value	0.000	0.883	0.069	0.714	0.000	0.875
PGI x Middle Tercile	0.107	0.136	-0.014	-0.025	0.242	0.171
S.E.	0.067	0.084	0.009	0.013	0.065	0.087
p-value	0.117	0.105	0.121	0.042	0.000	0.049
PGI x Top Tercile	0.065	0.007	-0.006	-0.028	0.271	0.151
S.E.	0.068	0.092	0.009	0.013	0.065	0.095
p-value	0.340	0.936	0.468	0.028	0.000	0.113
R^2 (Overall)	0.147	0.103	0.166	0.136	0.133	0.122
R^2 (Between)		0.129		0.126		0.138
R^2 (Within)		0.042		0.151		0.093
N	26612	26612	13102	13102	26898	27034
Sibling Groups		12933		6395		13136

This table shows Ordinary Least Squares (OLS) regression results for regressing each of the outcomes (Educational Attainment (EA), Imputed log hourly wages and Body Mass Index (BMI)), on their respective polygenic indices (PGI), infant mortality terciles and interactions. Infant mortality is reverse coded such that higher numbers are good. Columns 1 and 2 show results for EA, columns 3 and 4 for log hourly wage, and columns 5 and 6 for BMI. Columns 2, 4 and 6 include family fixed effects. All regressions included the following control variables: age, age squared, age cubed, gender, gender interacted with age variables, twenty principal components of the genetic data and dummies for genotyping batches. In the family fixed effects models some control variables had to be dropped due to multi-collinearity.

SI Table 5. Random Effects Models

	EA	Log Hourly Wage	BMI
	(1)	(2)	(3)
PGI	0.833	0.048	1.612
S.E.	0.048	0.005	0.046
p-value	0.000	0.000	0.000
Infant mortality rate	6.413	-0.148	-6.207
S.E.	7.675	1.088	7.972
p-value	0.403	0.892	0.436
PGI x Infant mortality rate	1.267	-0.693	-8.231
S.E.	5.474	0.740	5.218
p-value	0.817	0.349	0.115
R^2 (Overall)	0.147	0.157	0.132
R^2 (Between)	0.194	0.165	0.157
R^2 (Within)	0.037	0.145	0.090
N	26612	13102	27034
Sibling Groups	12933	6395	13136

This table shows the results of the random effects models based on a Mundlak formulation. Column (1) shows results for educational attainment (EA), column (2) for imputed log hourly wages, column (3) for body mass index (BMI). The outcomes were regressed on the PGI, infant mortality rate, their interaction and control variables (gender, year of birth and year of birth squared). Infant mortality rate is reverse coded such that higher numbers are good. For each variable within-family means were added to control for between family variation. See equation 2 for the full model.

7. References

Becker, J., Burik, C. A. P., Goldman, G., Wang, N., Jayashankar, H., & Karlsson Linnér, R. (2020). The Polygenic Index Repository: Resouce Profile and User Guide. *Manuscript in Submission*.

Berndt, S. I., Gustafsson, S., Mägi, R., Ganna, A., Wheeler, E., Feitosa, M. F., … Ingelsson, E. (2013). Genome-wide meta-analysis identifies 11 new loci for anthropometric traits and provides insights into genetic architecture. *Nature Genetics*, *45*(5), 501–512. https://doi.org/10.1038/ng.2606

Dupuis, J. J., Langenberg, C., Prokopenko, I., Saxena, R., Soranzo, N., Jackson, A. U., … Serrano-Rios, M. (2010). New genetic loci implicated in fasting glucose homeostasis and their impact on type 2 diabetes risk. *Nature Genetics*, *42*(2), 105–116. https://doi.org/10.1038/ng.520

Hodrick, R. J., & Prescott, E. C. (1997). Postwar U.S. Business Cycles: An Empirical Investigation. *Journal of Money, Credit and Banking*, *29*(1), 1–16. https://doi.org/10.2307/2953682

International HapMap 3 Consortium, T. I. H. 3, Altshuler, D. M., Gibbs, R. A., Peltonen, L., Altshuler, D. M., Gibbs, R. A., … McEwen, J. E. (2010). Integrating common and rare genetic variation in diverse human populations. *Nature*, *467*(7311), 52–58. https://doi.org/10.1038/nature09298

Kilpeläinen, T. O., Carli, J. F. M., Skowronski, A. A., Sun, Q., Kriebel, J., Feitosa, M. F., … Loos, R. J. F. (2016). Genome-wide meta-analysis uncovers novel loci influencing circulating leptin levels. *Nature Communications*, *7*, 10494. https://doi.org/10.1038/ncomms10494

Kweon, H., Burik, C. A. P., Karlsson Linnér, R., de Vlaming, R., Okbay, A., Martschenko, D., … Koellinger, P. D. (2020). *Genetic Fortune: Winning or Losing Education, Income, and Health* (Tinbergen Institute Discussion Papers No. 20- 053/V). Amsterdam.

Lee, J. J., Wedow, R., Okbay, A., Kong, E., Maghzian, O., Zacher, M., … others. (2018). Gene discovery and polygenic prediction from a 1.1-million-person GWAS of educational attainment. *Nature Genetics*, *50*(8), 1112.

Locke, A. E., Kahali, B., Berndt, S. I., Justice, A. E., Pers, T. H., Day, F. R., … Speliotes, E. K. (2015). Genetic studies of body mass index yield new insights for obesity biology. *Nature*, *518*(7538), 197–206. https://doi.org/10.1038/nature14177

Loh, P.-R., Tucker, G., Bulik-Sullivan, B. K., Vilhjálmsson, B. J., Finucane, H. K., Salem, R. M., … Price, A. L. (2015). Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nature Genetics*, *47*(3), 284–290. https://doi.org/10.1038/ng.3190

Lu, Y., Day, F. R., Gustafsson, S., Buchkovich, M. L., Na, J., Bataille, V., … Loos, R. J. F. (2016). New loci for body fat percentage reveal link between adiposity and cardiometabolic disease risk. *Nature Communications*, *7*, 10495. https://doi.org/10.1038/ncomms10495

McCarthy, S., Das, S., Kretzschmar, W., Delaneau, O., Wood, A. R., Teumer, A., … Haplotype Reference Consortium. (2016). A reference panel of 64,976 haplotypes for genotype imputation. *Nature Genetics*, *48*(10), 1279–1283. https://doi.org/10.1038/ng.3643

Rietveld, C. A. C. A., Medland, S. E. S. E., Derringer, J., Yang, J., Esko, T., Martin, N. G. N. W. N. W. N. G., … Koellinger, P. D. P. D. (2013). GWAS of 126,559 individuals identifies genetic variants associated with educational attainment. *Science*, *340*(6139), 1467–1471. https://doi.org/10.1126/science.1235488

Shungin, D., Winkler, T. W., Croteau-Chonka, D. C., Ferreira, T., Locke, A. E., Mägi, R., … Mohlke, K. L. (2015). New genetic loci link adipose and insulin biology to body fat distribution. *Nature*, *518*(7538), 187–196. https://doi.org/10.1038/nature14132

Trampush, J. W., Yang, M. L. Z., Yu, J., Knowles, E., Davies, G., Liewald, D. C., … Lencz, T. (2017). GWAS meta-analysis reveals novel loci and genetic correlates for general cognitive function: a report from the COGENT consortium. *Molecular Psychiatry*, *22*(3), 336–345. https://doi.org/10.1038/mp.2016.244

Turley, P., Walters, R. K., Maghzian, O., Okbay, A., Lee, J. J., Fontana, M. A., … Benjamin, D. J. (2018). Multi-trait analysis of genome-wide association summary statistics using MTAG. *Nature Genetics*, *50*(2), 229–237. https://doi.org/10.1038/s41588-017-0009-4

Vilhjálmsson, B. J., Jian Yang, H. K. F., Alexander Gusev, S. L., Stephan Ripke, G. G., Po-Ru Loh, Gaurav Bhatia, R. Do, Tristan Hayeck, H.-H. W., … Price, A. L. (2015). Modeling Linkage Disequilibrium Increases Accuracy of Polygenic Risk Scores. *The American Journal of Human Genetics*, *97*(4), 576–592. https://doi.org/10.1016/j.ajhg.2015.09.001

Winkler, T. W., Day, F. R., Croteau-Chonka, D. C., Wood, A. R., Locke, A. E., Mägi, R., … Loos, R. J. F. (2014). Quality control and conduct of genome-wide association meta-analyses. *Nature Protocols*, *9*(5), 1192–1212.

Zheng, J., Erzurumluoglu, A. M., Elsworth, B. L., Kemp, J. P., Howe, L., Haycock, P. C., … Neale, B. M. (2017). LD Hub: a centralized database and web interface to perform LD score regression that maximizes the potential of summary level GWAS data for SNP heritability and genetic correlation analysis. *Bioinformatics*, *33*(2), 272–279. https://doi.org/10.1093/bioinformatics/btw613